Antoine Caubrière, Sahar Ghannay, Natalia Tomashenko,
Renato De Mori, Antoine Laurent, Emmanuel Morin, Yannick Estève

ICASSP - May 2020

Error Analysis Applied to End-to-End Spoken Language Understanding

Introduction

Context

Analysis our End-to-End (E2E) Spoken Language Understanding (SLU) system

This system reaches state-of-the-art performance for a french SLU task

Introduction

Context

Analysis our End-to-End (E2E) Spoken Language Understanding (SLU) system

This system reaches state-of-the-art performance for a french SLU task

Goal

Analyze the errors produced by the system

Understand the weakness of this E2E system

From the weakness, discover how to improve our approach

Analysed system

Deep Speech 2 (DS2) **[Amodei *et al.*] (2016)**

- End-to-end speech recognition system
- Connectionist Temporal Classification (CTC)
- Allow the system to learn the alignment between speech and output sequence to produce

Analysed system

Deep Speech 2 (DS2) **[Amodei *et al.*] (2016)**

End-to-end speech recognition system

Connectionist Temporal Classification (CTC)

Allow the system to learn the alignment between speech and output sequence to produce

End-to-End Spoken Language Understanding (SLU) **[Ghannay *et al.*] (2018)**

Tag's boundaries injection

ASR : The sculptor Caesar died yesterday in Paris at the age of seventy-seven years

NER : The sculptor **<pers** Caesar **>** died **<time** yesterday **>** in **<loc** Paris **>** at the age of **<amount** seventy-seven years **>**

Analysed system

Deep Speech 2 (DS2) [Amodei *et al.*] (2016)

End-to-end speech recognition system

Connectionist Temporal Classification (CTC)

Allow the system to learn the alignment between speech and output sequence to produce

End-to-End Spoken Language Understanding (SLU) [Ghannay *et al.*] (2018)

Tag's boundaries injection

ASR : The sculptor Caesar died yesterday in Paris at the age of seventy-seven years

NER : The sculptor **<pers** Caesar **>** died **<time** yesterday **>** in **<loc** Paris **>** at the age of **<amount** seventy-seven years **>**

Curriculum-based transfer learning (CTL) [Caubrière *et al.*] (2019)

Train the same model through a sequence of training processes and transfer learning

Keep all parameters except the top layer

Use of different tasks sorted from the most generic to the most specific

Learned tasks

Automatic Speech Recognition (ASR)

Learned tasks

Automatic Speech Recognition (ASR)

Named Entity Recognition (NER)

Annotation according to 8 entity-types (*pers*, *loc*, *amount*, *etc*)

Learned tasks

Automatic Speech Recognition (ASR)

Named Entity Recognition (NER)

Annotation according to 8 entity-types (*pers*, *loc*, *amount*, *etc*)

Merged semantic concepts extraction (SC_mer)

MEDIA: French hotel booking task

PORTMEDIA: French theater ticket booking task

Annotation according to 76 semantic concepts (*location-town*, *stay-nbNight*, *nb-reservation*, *etc*)

Learned tasks

Automatic Speech Recognition (ASR)

Named Entity Recognition (NER)

Annotation according to 8 entity-types (*pers*, *loc*, *amount*, *etc)*

Merged semantic concepts extraction (SC_mer)

MEDIA: French hotel booking task

PORTMEDIA: French theater ticket booking task

Annotation according to 76 semantic concepts (*location-town*, *stay-nbNight*, *nb-reservation*, *etc*)

Semantic concepts extraction on MEDIA (M)

Our target task

Learned tasks

Automatic Speech Recognition (ASR)

Named Entity Recognition (NER)

Annotation according to 8 entity-types (*pers*, *loc*, *amount*, *etc*)

Merged semantic concepts extraction (SC_mer)

MEDIA: French hotel booking task

PORTMEDIA: French theater ticket booking task

Annotation according to 76 semantic concepts (*location-town*, *stay-nbNight*, *nb-reservation*, *etc*)

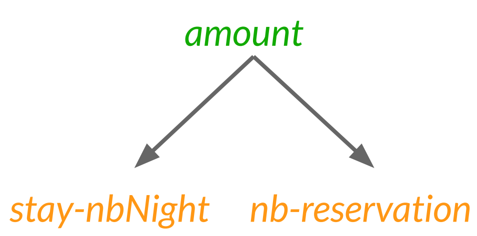
Semantic concepts extraction on MEDIA (M)

Our target task

Order of learned tasks

We define the following order of specificity:

Speech > Named Entities > Semantic Concepts



Data

French data sets

Uses of as much data as possible at our disposal

Broadcast news, Telephone and Human-Human Dialogue

Data

French data sets

Uses of as much data as possible at our disposal

Broadcast news, Telephone and Human-Human Dialogue

Speech ~ 360h

Data

French data sets

Uses of as much data as possible at our disposal

Broadcast news, Telephone and Human-Human Dialogue

Speech ~ 360h

NE ~ 300h

Data

French data sets

Uses of as much data as possible at our disposal

Broadcast news, Telephone and Human-Human Dialogue

Speech ~ 360h

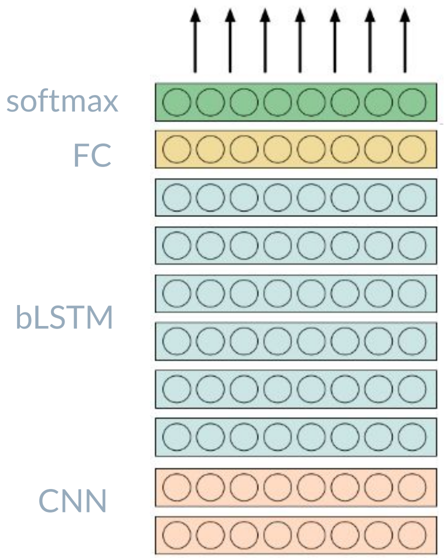
NE ~ 300h

SC ~ 25h

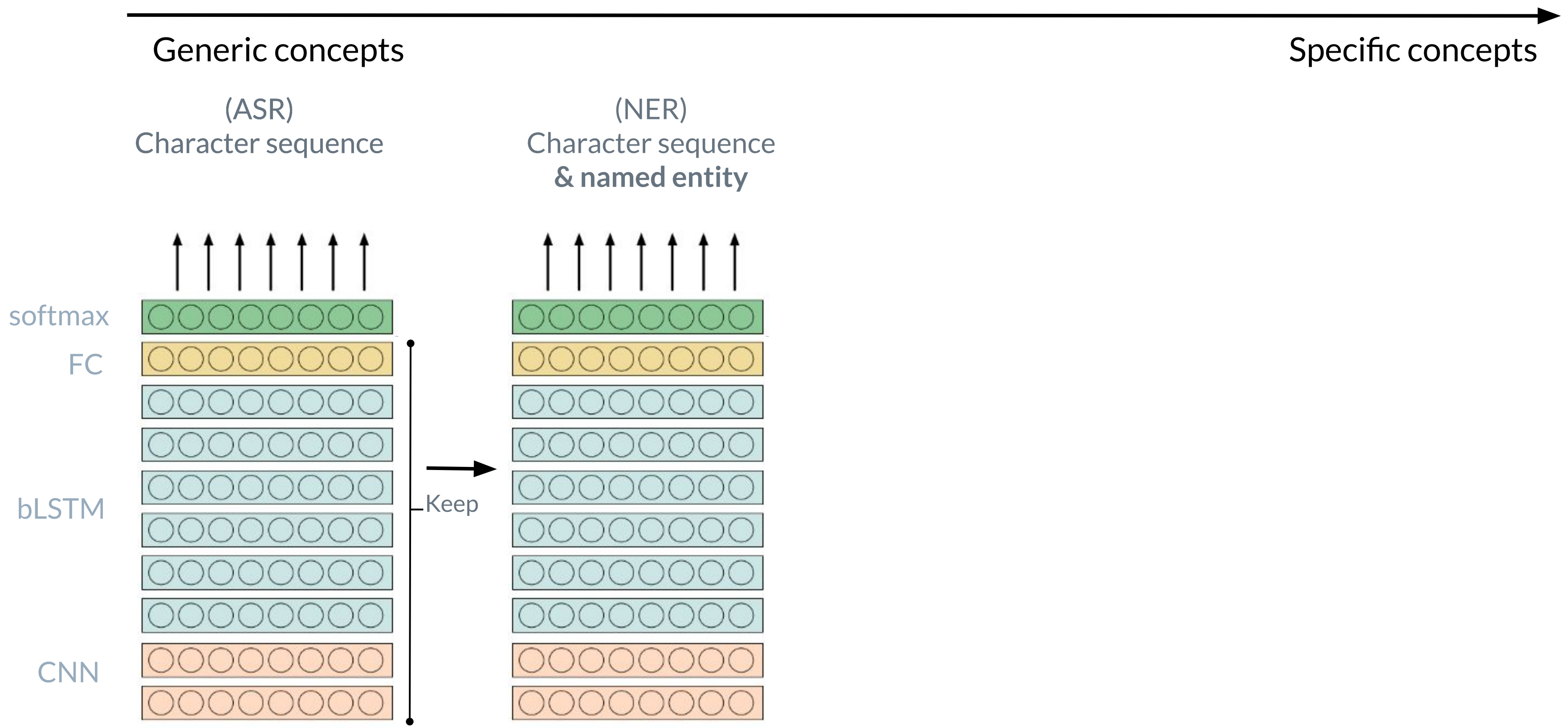
CTL approach

Generic concepts — Specific concepts

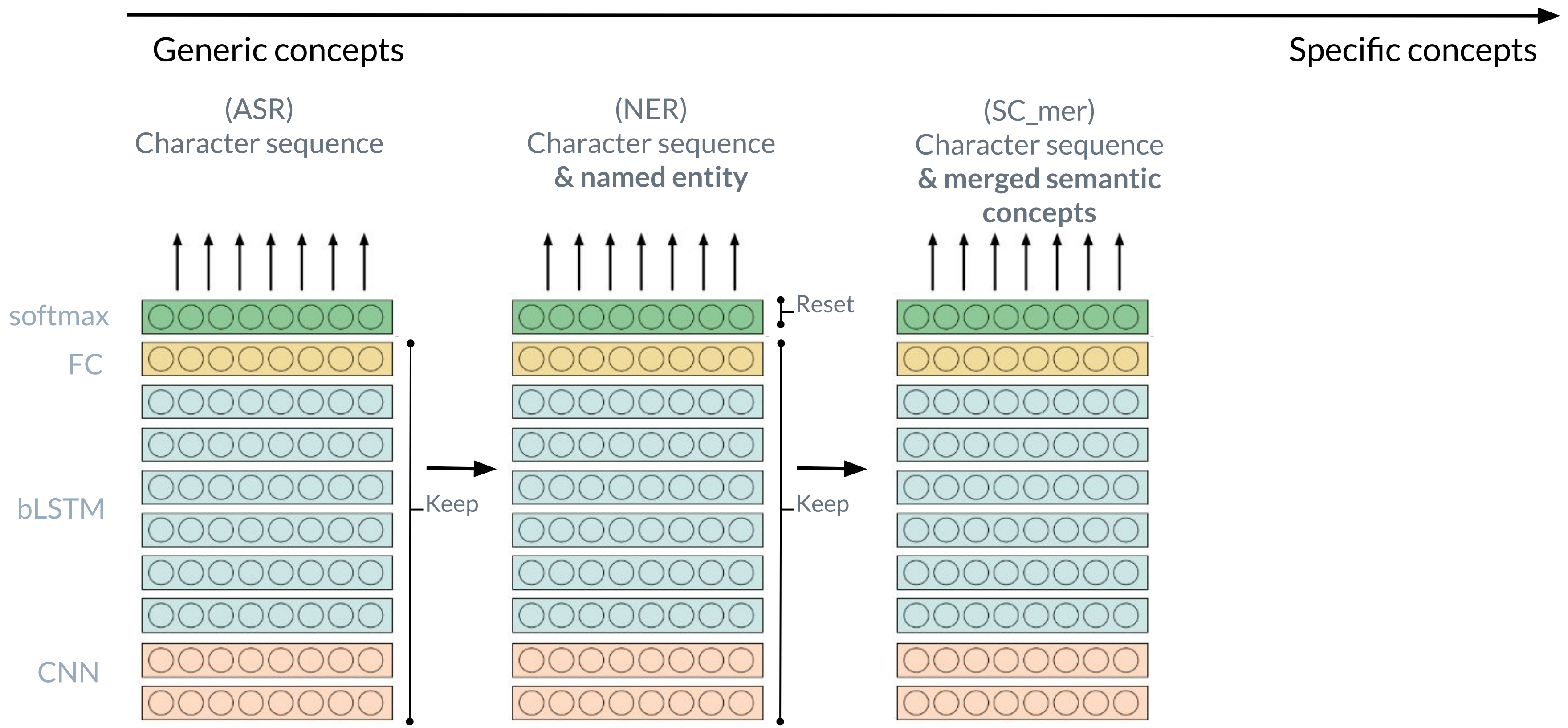
(ASR)
Character sequence



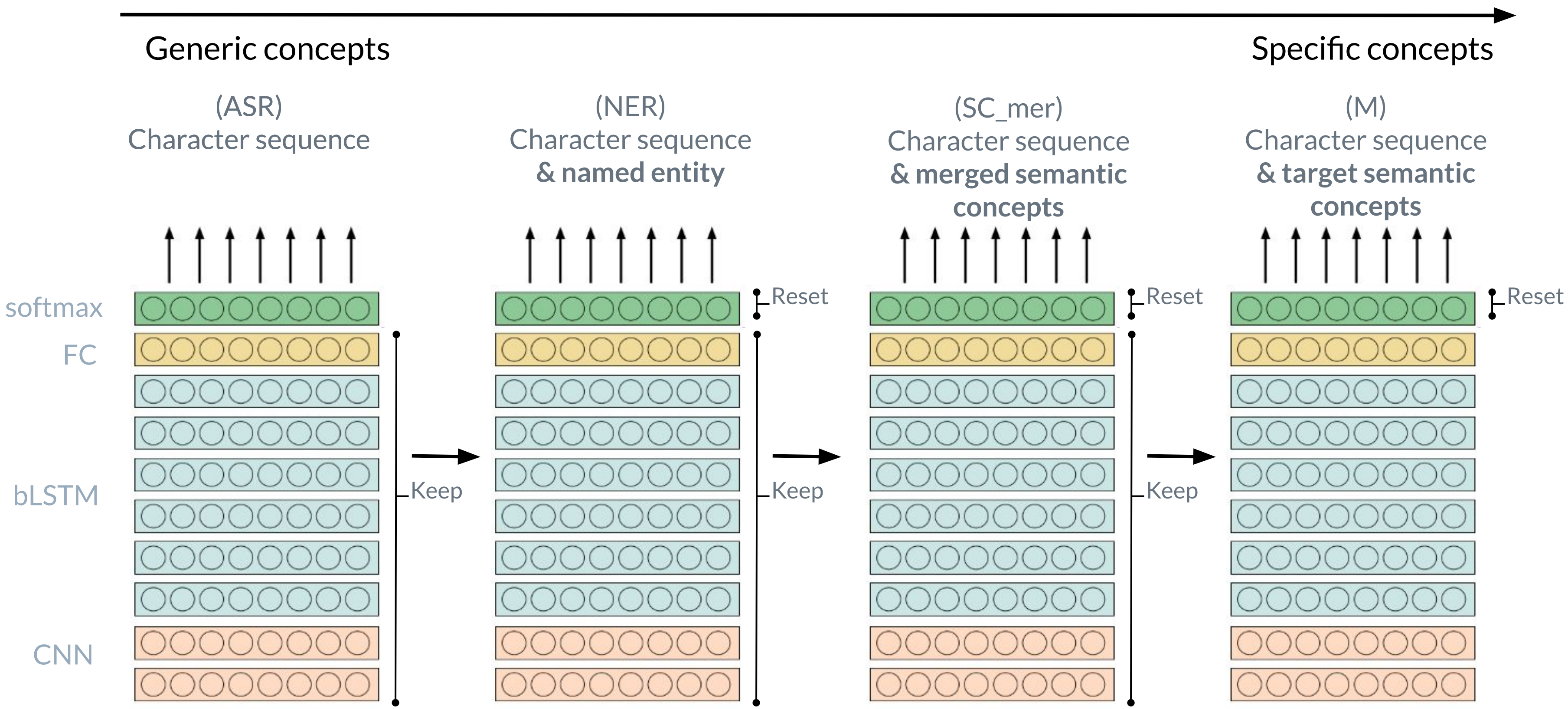
CTL approach



CTL approach



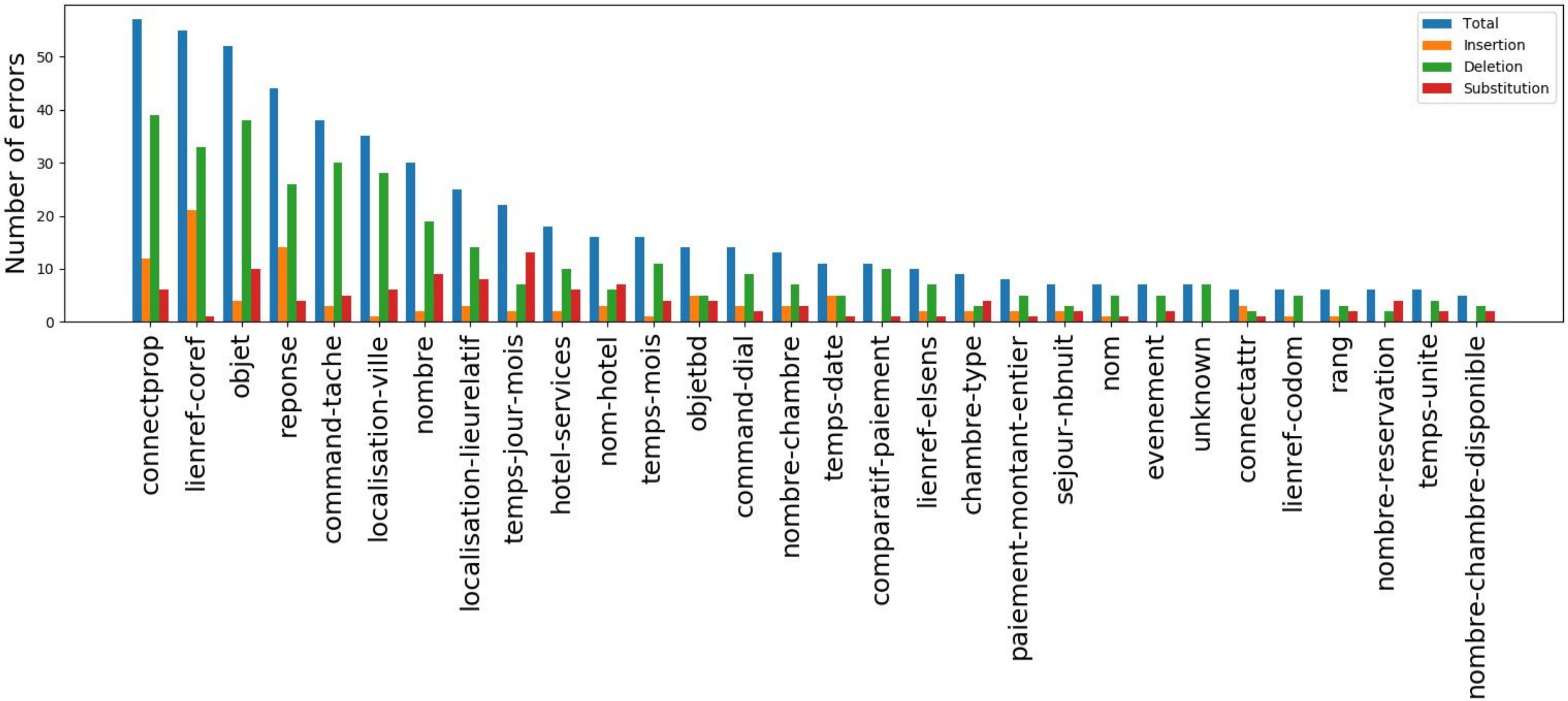
CTL approach



Errors distribution

Systems outputs for MEDIA concepts (development dataset)

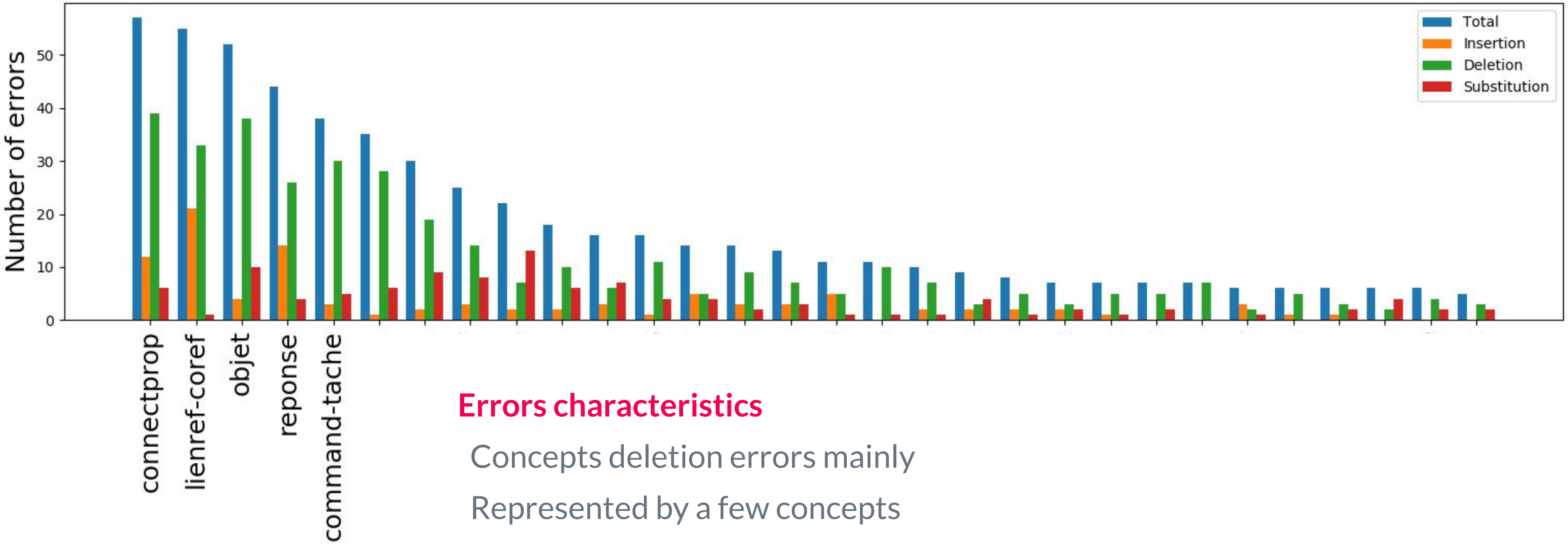
The thirty most common mistakes



Errors distribution

Systems outputs for MEDIA concepts (development dataset)

The thirty most common mistakes



Errors characteristics

Concepts deletion errors mainly

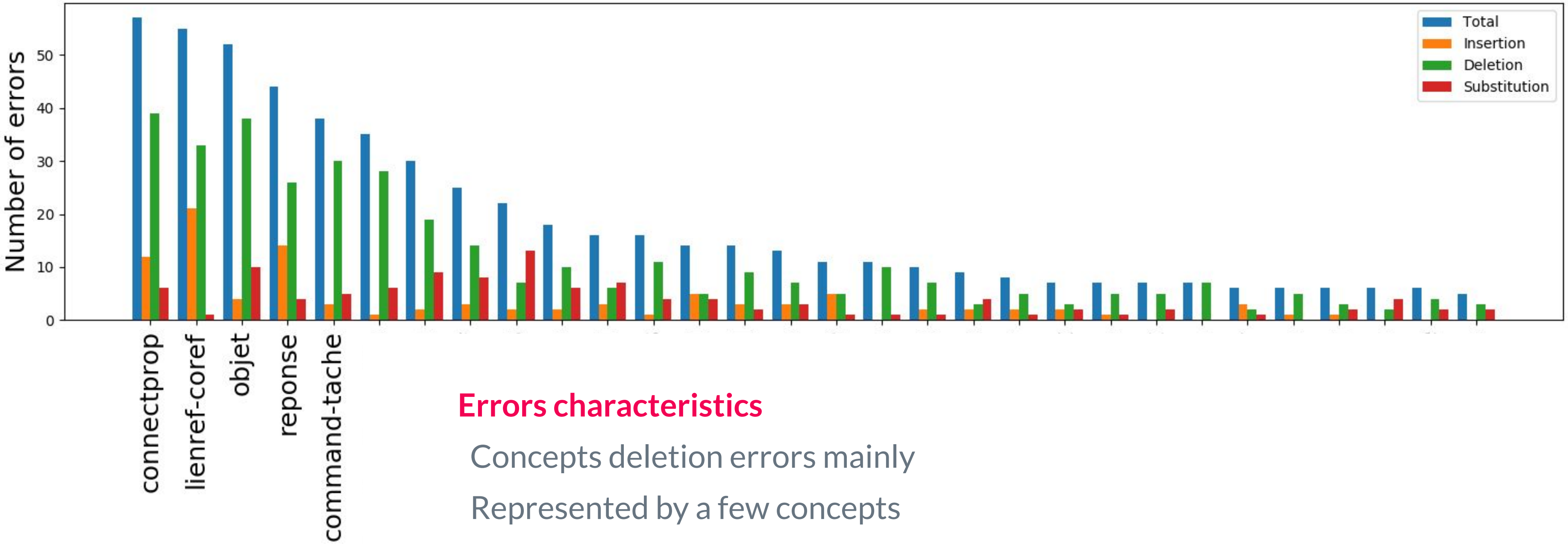
Represented by a few concepts

Frequent errors corresponding to concepts with small values (*connectprop*, *lienref-coref*, ...)

Errors distribution

Systems outputs for MEDIA concepts (development dataset)

The thirty most common mistakes



Errors characteristics

Concepts deletion errors mainly

Represented by a few concepts

Frequent errors corresponding to concepts supported by small words (*connectprop*, *lienref-coref*, ...)

-> "I'd like to know ***<lienref-coref*** the **>** ***<object*** price **>** of the night ***<connectprop*** and **>** if there are any ***<object*** rooms **>** left"

Transcription problem

Cases of concepts deletions (MEDIA development dataset)

Transcription problem

Cases of concepts deletions (MEDIA development dataset)

Reference of an example

-> “From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october **<connectprop** and **>** in **<location-town** Périgueux **>**”

Transcription problem

Cases of concepts deletions (MEDIA development dataset)

Reference of an example

-> "From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october **<connectprop** and **>** in **<location-town** Périgueux **>**"

Correct automatic transcription

-> "From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october and in **<location-town** Périgueux **>**"

Transcription problem

Cases of concepts deletions (MEDIA development dataset)

Reference of an example

-> "From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october **<connectprop** and **>** in **<location-town** Périgueux **>**"

Correct automatic transcription

-> "From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october and in **<location-town** Périgueux **>**"

Incorrect automatic transcription

-> "From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october **<connectprop** and **>** ***par lieu*** "

Transcription problem

Cases of concepts deletions (MEDIA development dataset)

Reference of an example

-> “From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october **<connectprop** and **>** in **<location-town** Périgueux **>**”

Correct automatic transcription

-> “From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october and in **<location-town** Périgueux **>**”

Incorrect automatic transcription

-> “From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october **<connectprop** and **>** ***par lieu*** ”

Correct automatic transcription but the value is nested in another concept

-> “From **<time-date** nineteen **>** to **<time-date** twenty-two **>** october **<location-town** and in Périgueux **>**”

Transcription problem

Focused concept	Nb Deletion	Correct ASR	Wrong ASR	Nested
connectProp	39	**28**	6	5
lienref-coref	33	**19**	10	4
objet	38	**31**	4	3

Transcription problem

Focused concept	Nb Deletion	Correct ASR	Wrong ASR	Nested
connectProp	39	**28**	6	5
lienref-coref	33	**19**	10	4
objet	38	**31**	4	3

Extra observation

Regularly ended tags without any associated started tags

-> "I'd like to know the **>** ***<object*** price **>** of the night"

A concept segmentation issue

Segmentation problem

Tackle the concept segmentation issue

Split the final MEDIA task into two tasks in the CTL approach

Firstly, train the system to retrieve the boundaries of concepts only (M_seg)

Secondly, train the system to specify the concepts (classical M task)

CTL approach become : ASR -> NER -> SC_mer -> *M_seg* -> M

Segmentation problem

Tackle the concept segmentation issue

Split the final MEDIA task into two tasks in the CTL approach

Firstly, train the system to retrieve the boundaries of concepts only (M_seg)

Secondly, train the system to specify the concepts (classical M task)

CTL approach become : ASR -> NER -> SC_mer -> *M_seg* -> M

Outputs to produce for M_seg

Replace each starting tag by a generic '<'

M : "I'd like to know ***<lienref-coref*** the **>** ***<object*** price **>** of the night ***<connectprop*** and **>** if there are any ***<object*** rooms **>** left"

M_seg : "I'd like to know **<** the **>** **<** price **>** of the night **<** and **>** if there are any **<** rooms **>** left"

Segmentation problem

Concept Error Rate (CER)

Evaluates concepts only

Concept Value Error Rate (CVER)

Evaluates concepts and values (Words within the concepts)

Segmentation problem

Concept Error Rate (CER)

Evaluates concepts only

Concept Value Error Rate (CVER)

Evaluates concepts and values (Words within the concepts)

System	CER*	CVER*
ASR -> NER -> SC_mer -> M	21.6	27.7
ASR -> NER -> SC_mer -> *M_seg* -> M	**20.7**	**27.2**
Relative gain	***+4.1%***	***+1.0%***

* Scores for MEDIA test set

ASR : Automatic Speech Recognition

NER : Named Entity Recognition

SC_mer : Merged Semantic Concept extraction

M_seg : MEDIA segmentation task

M : MEDIA task

NER task contribution

Unseen Concept/Value pairs (UCV)

Examples seen in the development dataset which do not appear in the training dataset

A total of 533 UCV

NER task contribution

Unseen Concept/Value pairs (UCV)

Examples seen in the development dataset which do not appear in the training dataset

A total of 533 UCV

System	Correct Concept/Value	Correct Value
ASR -> NER -> SC_mer -> M	132	38
ASR -> SC_mer -> M	124	36

NER task contribution

Unseen Concept/Value pairs (UCV)

Examples seen in the development dataset which do not appear in the training dataset

A total of 533 UCV

System	Correct Concept/Value	Correct Value
ASR -> NER -> SC_mer -> M	132	38
ASR -> SC_mer -> M	124	36

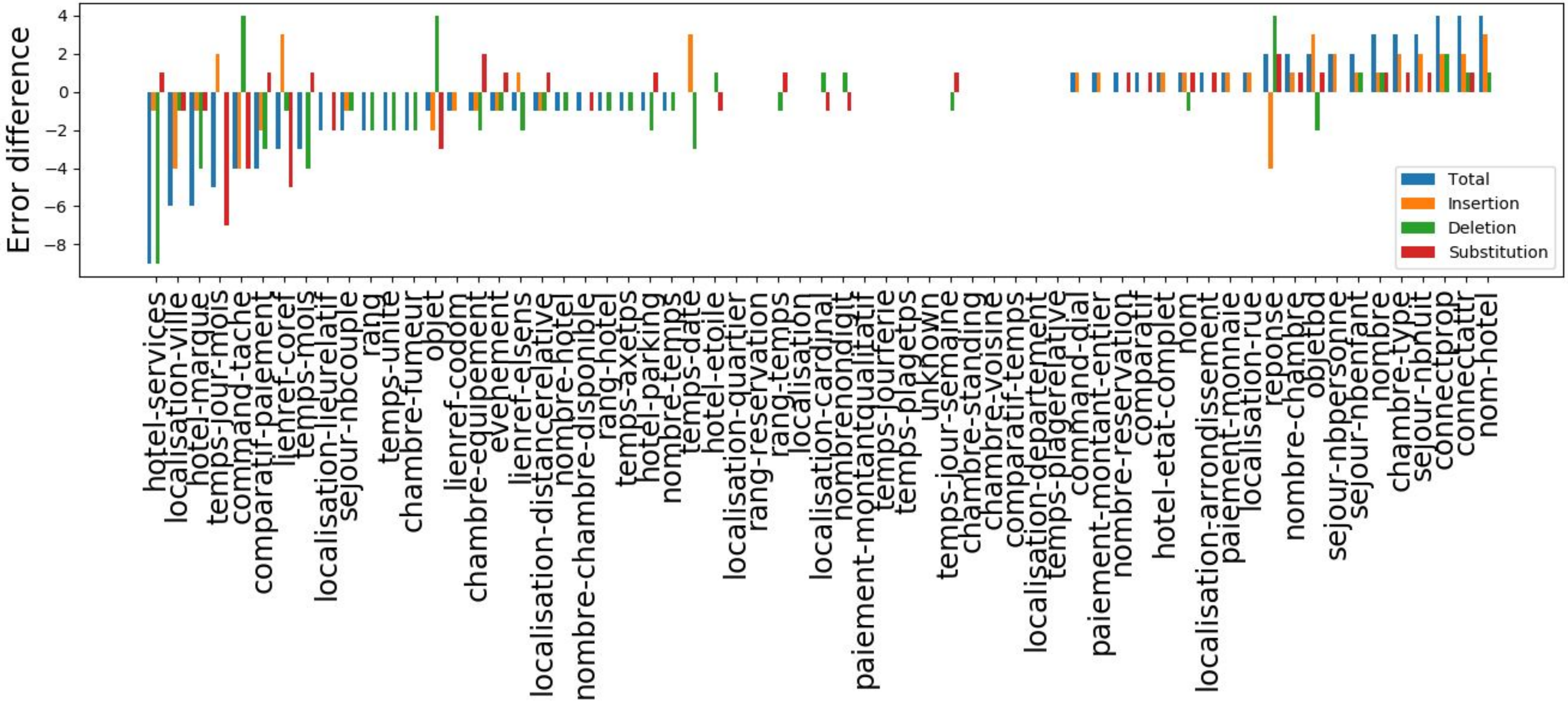
Only a small number of correct values (around a quarter)

Incorrect speech transcription for a major part of the UCV

NER task contribution

Delta of concepts recognition errors

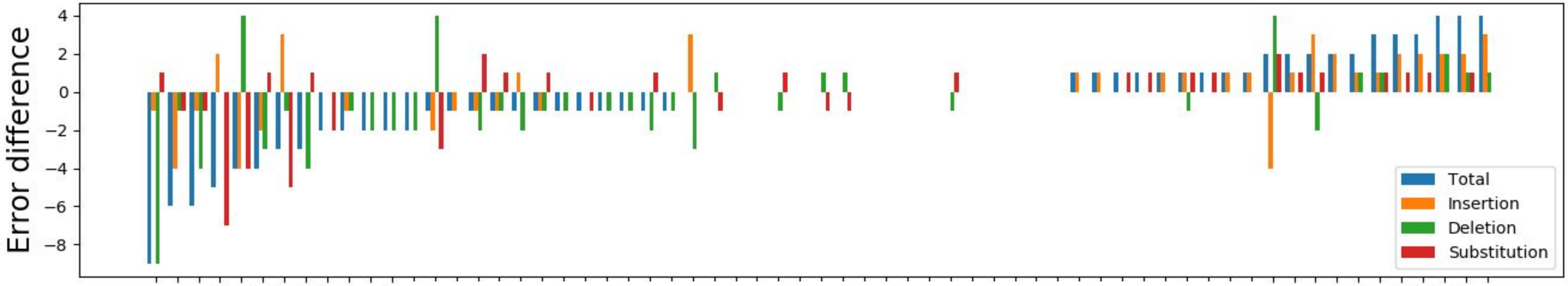
With and without the NER task during the training



NER task contribution

Delta of concepts recognition errors

With and without the NER task during the training



Observations

Strongly positive evolution

Improved concepts close to named entities

'Location-town', *'time-day-month'*, *'stay-nbcouple'* ...

Embeddings visualisation

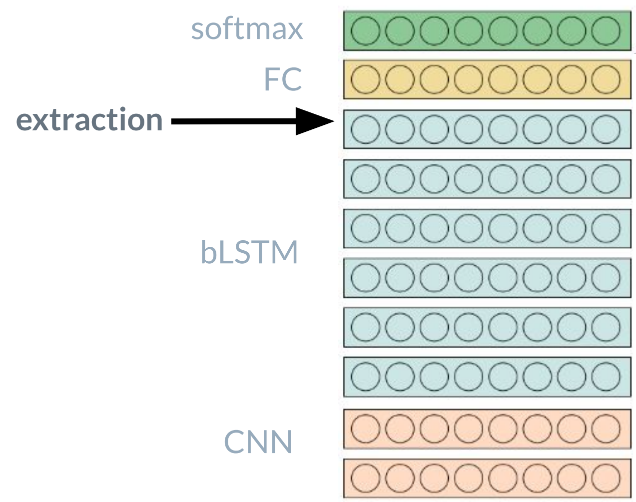
Embeddings extraction (MEDIA development dataset)

From the last bLSTM layer of DS2

DS2 trained with CTC loss function

One embedding per input frame

One character per input frame



Embeddings visualisation

Embeddings extraction (MEDIA development dataset)

From the last bLSTM layer of DS2

DS2 trained with CTC loss function

One embedding per input frame

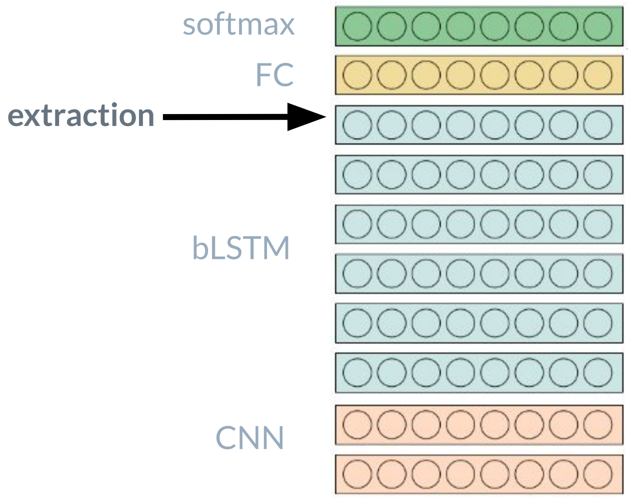
One character per input frame

Words and concepts representation

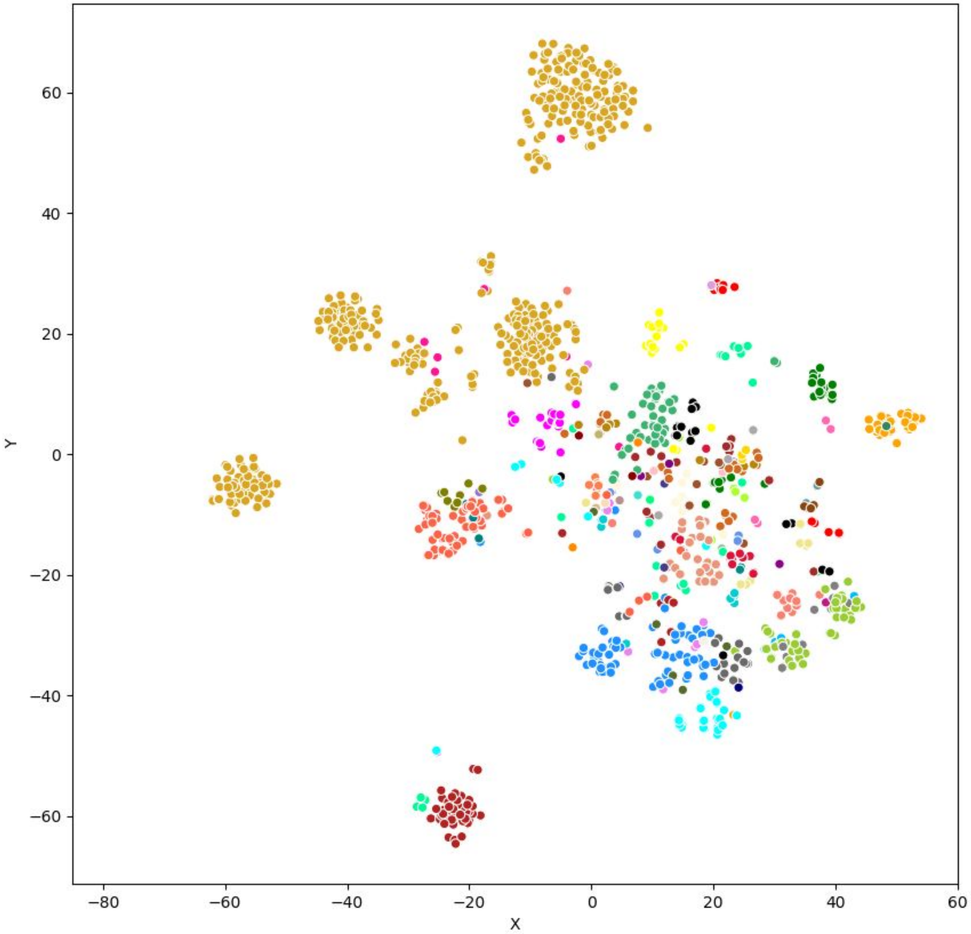
Represented by more than one character

Use of the sum of each frame's embeddings

Use of a t-SNE transformation for a 2D representation



Embeddings visualisation



Observation

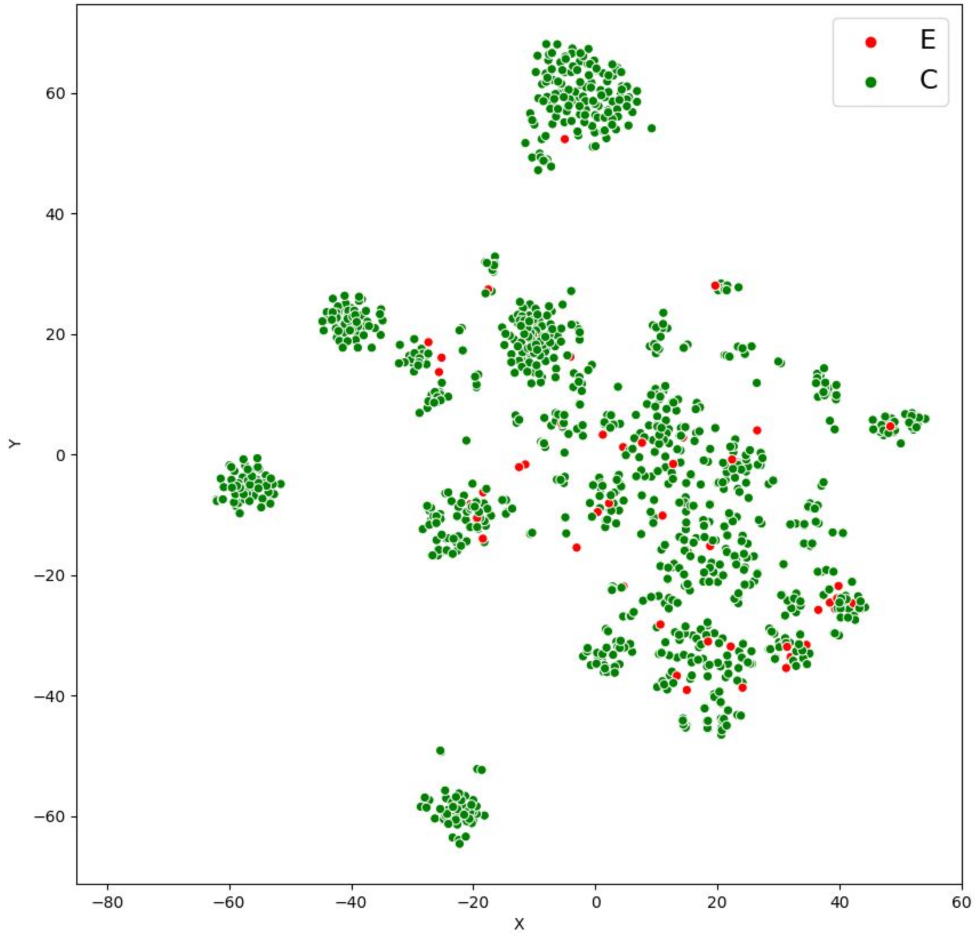
Each color represent a semantic class

Concepts of the same classes are clustered

Some very clear clusters

An area with mixed concepts

Embeddings visualisation



Green : well-recognized concepts

Red : badly-recognized concepts

Observation

Main errors are in the mixed area

Concepts errors seem to be related to an insufficiently discriminative internal representation

Conclusion

We presented a qualitative study of errors produced by an end-to-end SLU system

Conclusion

We presented a qualitative study of errors produced by an end-to-end SLU system

We observed that most of the errors concern generic and domain-independent concepts

Conclusion

We presented a qualitative study of errors produced by an end-to-end SLU system

We observed that most of the errors concern generic and domain-independent concepts

We detected a concept segmentation issue for our SLU system

Conclusion

We presented a qualitative study of errors produced by an end-to-end SLU system

We observed that most of the errors concern generic and domain-independent concepts

We detected a concept segmentation issue for our SLU system

We proposed an intermediate segmentation training task which allows 4.1% relative gain

Conclusion

We presented a qualitative study of errors produced by an end-to-end SLU system

We observed that most of the errors concern generic and domain-independent concepts

We detected a concept segmentation issue for our SLU system

We proposed an intermediate segmentation training task which allows 4.1% relative gain

We proposed a way to compute embeddings of sub-sequences

Conclusion

We presented a qualitative study of errors produced by an end-to-end SLU system

We observed that most of the errors concern generic and domain-independent concepts

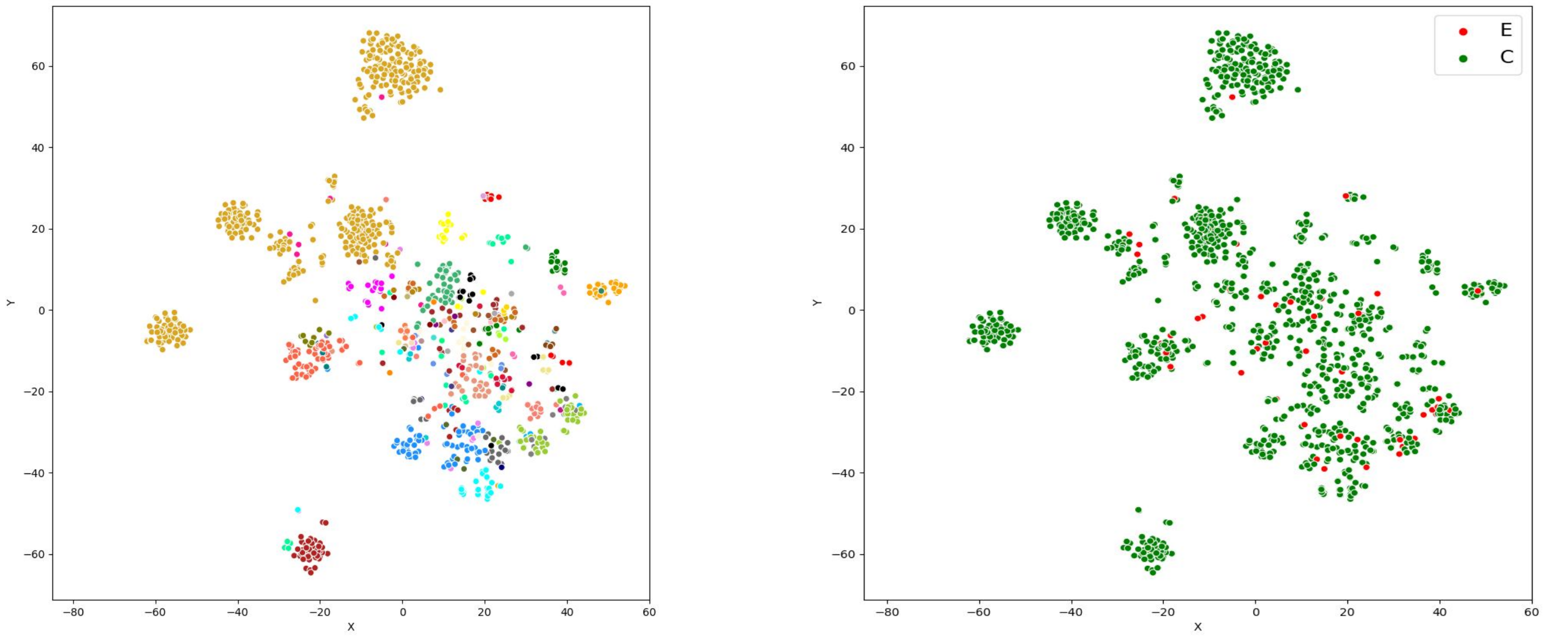
We detected a concept segmentation issue for our SLU system

We proposed an intermediate segmentation training task which allows 4.1% relative gain

We proposed a way to compute embeddings of sub-sequences

We observed that output concept errors appear to be related to an insufficiently discriminative internal representation

Perspectives



Take benefit from this cartography

How to take benefit to improve performances?
How to force the system to represent the concepts in a more relevant space?
Exploit the position of embeddings in the continuous space

Thank you

Contact: antoine.caubriere@univ-lemans.fr